Journal of the Korean Institute of Industrial Engineers
Vol. 38, No. 3, pp. 191-197, September 2012.

비모수 추정방법을 활용한 *k*NNDD의 이상치 탐지 기법

손정환 · 김성범†

고려대학교 산업경영공학부

*k*NNDD-based One-Class Classification by Nonparametric Density Estimation

Junghwan Son · Seoung Bum Kim

School of Industrial Management Engineering, Korea University

One-class classification (OCC) is one of the recent growing areas in data mining and pattern recognition. In the present study we examine a *k*-nearest neighbors data description (*k*NNDD) algorithm, one of the OCC algorithms widely used. In particular, we propose to use nonparametric estimation methods to determine the threshold of the *k*NNDD algorithm. A simulation study has been conducted to explore the characteristics of the proposed approach and compare it with the existing approach that determines the threshold. The results demonstrate the usefulness and flexibility of the proposed approach.

Keywords: Data Mining, One-Class Classification, *k*-Nearest Neighbor Data Description, Kernel Density Estimation, Bootstrap

1. 서 론

데이터마이닝(Data Mining)이란 대용량의 복잡한 데이터로부터 유용한 정보를 추출하고 이로부터 의미 있는 규칙이나 패턴을 찾는 과정을 의미한다(Berry and Linoff, 1997).

데이터마이닝을 이용하여 해결할 수 있는 문제는 크게 분류(classification), 예측(prediction), 그리고 군집(clustering)으로 나눌 수 있으며 응용 분야로는 의료, 제조, 서비스, 금융, 환경 등 다양하게 존재하고 있다(Hand *et al*., 2001).

분류와 예측은 종속변수에 대한 미래의 값을 예측한다는 점에서 공통된 목적을 가지고 있으나 종속변수가 범주형이면 분류, 연속형이면 예측으로 나눌 수 있다. 군집은 종속변수 없이 독립변수만의 특성을 이용해 비슷한 관측치끼리 그룹화하는 방법을 말한다.

본 논문에서 다루고자하는 데이터마이닝 기법은 단일 클래스 분류(One-class classification) 기법으로 이는 기존 지도형 분류(Supervised classification) 기법과는 차이가 있다. 지도형 분류 기법은 범주형인 종속변수와 이를 설명하는 독립변수들과의 관계를 통해 분류모델을 구축하고 미래 관측치의 독립변수 값이 주어졌을 때 알려지지 않은 해당 종속변수의 값을 예측한다(Khan and Madden, 2010).

반면 단일 클래스 분류 기법은 범주형 종속변수의 미래의 값을 예측한다는 점에서는 기존 지도형 분류 기법과 동일하나 분류모델 생성 시 종속변수를 사용하지 않는다는 점에서 차이가 있다. 최근 단일 클래스 분류기법이 활발히 연구되고 있으며 그 응용분야로 정보검색, 얼굴 인식, 생물정보학, 저자판별 분류, 문서분류, 제조업체 모니터링 등 다양하게 존재하고 있다(Manevitz and Yousef, 2001; Sanchez-Yanez, 2003; Koppel and Schler, 2004; Kim *et al*., 2011).

단일 클래스 분류에서 모델을 구축하는 과정을 좀 더 자세히 설명하자면 학습데이터의 형태를 잘 아우를 수 있는 수 있는 폐쇄 경계선(closed boundary)을 결정하고, 이를 통하여 미래 관측치가 정상 데이터에 속하는지 아닌지를 분류한다. 즉, 미래의 관측치가 폐쇄 경계선 안에 있으면 기존 패턴을 따르는 데이터

본 연구는 한국연구재단의 기초연구 사업지원(0003811)과 지식경제부의 정보통신연구기반구축사업(NIPA-2011-(B1110-1101-0002))의 지원으로 수행된 연구임.
† 연락저자 : 김성범 교수, 136-701 서울시 성북구 안암동 5가 1번지 고려대학교 산업경영공학부, Tel : 02-3290-3397, Fax : 02-929-5888
E-mail : sbkim1@korea.ac.kr
2012년 2월 20일 접수; 2012년 6월 7일; 수정본 접수; 2012년 6월 12일 게재 확정.

로 경계선 밖에 있으면 새로운 패턴을 가지고 있는 데이터로 분류하는 것이다. 예를 들어 품질관리문제에서 제품 불량여부를 모니터링하는 문제에 적용해 보면 다음과 같다. 우선 양품 데이터만을 사용하여 단일 클래스 분류 모델을 구축하고 이를 이용하여 미래 제품의 관측치가 구축된 폐쇄 경계 안에 들어올 경우 정상으로 밖에 있을 경우에는 불량으로 판단하는 것이다.

단일 클래스 분류 기법은 일반적으로 분포를 이용한 추정법(density estimation), 경계선을 통한 분류 방법(boundary methods), 그리고 재구축 방법(reconstruction methods)으로 구분하고 있다. 그 중에서 경계선을 통한 분류 기법이 많이 사용되고 있는데 이는 데이터의 확률분포에 대한 가정이 필요 없고 변수의 특성에 관계없이 비교적 정확한 성능을 보이고 있기 때문이다. 경계선을 통한 분류 방법으로는 *k*-centers, support vector data description, *k*-nearest neighbors data description(*k*NNDD) 등의 방법이 있다(Tax, 2001).

경계선을 기반으로 하는 단일 클래스 분류 기법들의 기본적인 아이디어는 새로운 관측치 가 주어졌을 때 이를 정상 관측치들과 거리를 계산하고 그 거리가 정해진 임계값보다 큰지 작은지를 판단하여 정상과 이상으로 분류하는 것이다. 따라서 경계선 기반 단일 클래스 기법은 거리종류를 정하는 문제와 임계값을 결정하는 두 문제로 세분화 할 수 있다.

본 연구에서는 경계선 기반 단일 클래스 분류 기법의 대표적인 방법인 *k*NNDD기법에서 현재 사용되고 있는 임계값 결정 방식의 한계를 보이고 이를 개선할 수 있는 방법을 제안하고자 한다. 본 연구에서는 제안하고자 하는 임계값 결정 방법은 비모수 추정 기법인 kernel density estimation(KDE) 기법과 붓스트랩(bootstrap) 기법을 기반으로 하고 있다.

본 논문의 구성은 다음과 같다. 제 2장에서는 kNNDD 알고리즘에 대해 설명하고 제 3장에서는 제안하는 비모수 추정 기반 임계값 결정 과정에 관해 설명하도록 하겠다. 제 4장에서는 제안하는 방법의 특성과 유용성을 보이기 위해 시뮬레이션을 실시하고 그 결과를 설명하도록 하겠다. 마지막으로 제 5장에서는 제안한 방법에 대한 전체적인 요약을 하도록 하겠다.

2. *k*-Nearest Neighbors Data Description(*k*NNDD)

이번 장에서는 본 연구에서 중점적으로 다루고자 하는 *k*NNDD 알고리즘에 대해 설명을 하도록 하겠다. *k*NNDD는 기본적으로 최인접 분류기(nearest neighbor classifier)를 기반으로 하고 있는 단일 클래스 분류 기법이다(Duda and Hart, 1973).

먼저 <Figure 1>을 통해 kNNDD 알고리즘에 대한 개괄적인 설명을 하도록 하겠다. 그림에서 d_1은 새로운 관측치 x와 인접한 k_1개의 이웃 관측치 사이의 평균 거리를 나타내며, d_2는 k_1개의 이웃 관측치들과 인접한 k_2개의 이웃 관측치 사이의 평균거리를 나타내고 있다. 새로운 관측치 x가 이상인지 여부는 d_1과 d_2를 비교하여 $\frac{d_1}{d_1} > 1$이면 이상 관측치로 판단한다. d_1과 d_2의 개념을 아래 식들을 통해 좀 더 자세히 설명해 보도록 하겠다.

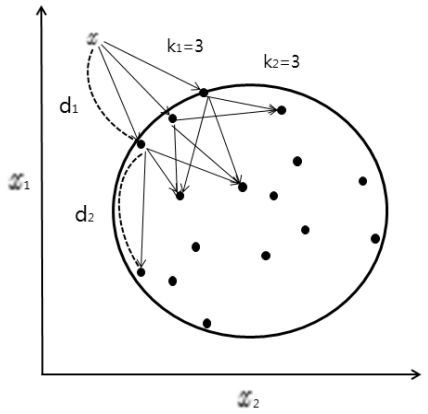


Figure 1. The basic concept of a *k*nndd algorithm

식 (1)에서 $\| x \|$는 유클리디언 거리(Euclidean distance)를 말하며 $NN_i(x)$는 새로운 관측치 x와 i번째 가까운 이웃 관측치를 나타낸다.

$$d_1(x) = \sum_{i=1}^{k_1} \frac{\| x - NN_i(x) \|}{k_1} \tag{1}$$

즉, $d_1(x)$는 관측치 x와 인접한 k_1개의 이웃 관측치들 사이의 평균거리를 의미한다.

$d_2(x)$는 위에서 결정한 x의 이웃 관측치들과 그것들의 이웃 관측치들 사이의 평균거리이고 계산식은 다음과 같다.

$$d_2(x) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{\| NN_i(x) - NN_j(NN_i(x)) \|}{k_1 \cdot k_2} \tag{2}$$

즉, $d_2(x)$는 x와 인접한 k_1개의 이웃 관측치들 구하고 이것들 각각과 인접한 k_2개의 이웃 관측치들을 찾아 평균거리를 구한 값이다.

*k*NNDD 알고리즘은 위에서 설명한 $d_1(x)$와 $d_2(x)$의 비를 통해 거리를 정의하며 이는 식 (3)과 같이 나타낼 수 있다.

$$\frac{d_1(x)}{d_2(x)} = \frac{k_2 \sum_{i=1}^{k_1} \| x - NN_i(x) \|}{\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \| NN_i(x) - NN_j(NN_i(x)) \|} \tag{3}$$

*k*NNDD의 거리는 <Figure 1>을 통해 그 의미를 살펴 볼 수 있는데 즉 $d_1(x)$의 거리가 크면 클수록 관측치 x는 이상치일 확률이 커짐을 알 수 있다. 하지만 $d_1(x)$를 사용한 거리는 그것의 크고 작음에 대한 기준이 없기 때문에 $d_2(x)$를 사용해 상대적인 거리를 나타낼 수 있도록 한다. 따라서 $d_1(x)$과 $d_2(x)$의 비는 두 거리가 같을 때면 1이 되고, $d_1(x)$가 크면

1보다 큰 수가 될 것이며, 반대일 경우에는 0과 1사이의 수가 될 것이다. 여기서 $d_1(x)$와 $d_2(x)$의 비가 얼마나 커야 이상치로 판단할 수 있는지에 대한 기준(임계값)이 필요한데 기존에는 임계값을 단순히 1로 정의하고 있으며 아래의 식으로 표현할 수 있다(Tax, 2001; Sukchotrat, 2009; Kim *et al*., 2011).

$$\frac{d_1(x)}{d_2(x)} > 1 \;\Rightarrow\; reject\ x \tag{4}$$

즉, $d_1(x)$가 $d_2(x)$보다 크면 이상 관측치로 작으면 정상 관측치로 분류한다.

하지만 모든 데이터에 임계값 1을 적용한다는 것은 데이터의 특성과 분포를 고려하지 않은 일률적인 방법으로 한계가 있을 수 있다. 따라서 본 논문에서는 이러한 기존 임계값 결정 방식의 문제점을 보여주고 이를 해결하기 위한 방법을 제안해 보도록 하겠다.

3. 비모수 추정 방법을 활용한 임계값 결정

비모수 추정 방법은 함수형태를 가정하지 않고 주어진 데이터로부터 직접 함수를 추정하는 방법이다(Wasserman, 2006).

실제 문제에서도 대부분의 데이터들은 분포가 알려져 있지 않고 다양한 형태로 존재하기 때문에 확률분포에 대한 사전 지식이 없을 경우 순수하게 주어진 데이터만을 통해 분포를 추정하는 비모수 추정 방법이 널리 사용되고 있다(Song *et al*., 2003).

비모수 추정 방법으로는 히스토그램, 가우시안(Gaussian), KDE, bootstrap 등 여러 가지 방법들이 있다(Silverman, 1986; Efron, 1993). 이 중 본 연구에서는 KDE와 bootstrap 기법을 사용하였다.

3.1 Kernel Density Estimation(KDE)를 이용한 임계값 결정

KDE는 비모수 추정 방법 중의 하나로 커널(kernel)을 사용하여 분포가 알려지지 않은 데이터의 분포를 추정하는 방법이다. $x_1, x_2, \cdots, x_N$이 서로 독립적이고 동일한 분포를 따르는 확률변수들의 표본이라면 KDE에 의해 추정된 함수식은 다음과 같다.

$$\hat{f}_h(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - kNNDD_i}{h}\right) \tag{4}$$

여기서 K는 커널을 나타내며 그 종류로는 균일분포(uniform), 정규분포(normal), 이차(quadratic), 삼각(triangular), 코사인(cosine), Epachenikov 커널 등이 있다. 본 연구에서는 가장 보편적으로 쓰이고 있는 정규분포 커널을 사용하였다(Martinez and Martinez, 2007). 또한 대역폭(bandwidth) h는 평활 모수(smoothing parameter)로써 그 값이 커질수록 분포가 보다 평활하게 추정되어진다(Chou *et al*., 2001). 본 논문에서는 Sheather and Jones(1991)가 제안하고 있는 방법을 통해 대역폭을 결정하였다.

KDE를 통해 임계값을 결정하는 방법은 추정된 $\hat{f}_h(x)$을 이용하여 $100 \cdot (1-\alpha)$분위수를 사용하는 것이다. 여기서 α는 0에서 1사이의 값이며 사용자가 임의로 결정할 수 있다. 이를 식으로 표현하면 다음과 같다.

$$\text{임계값} = \hat{f}(x)^{-1}(1-\alpha) \tag{5}$$

위의 임계값을 구하기 위해서는 $\hat{f}_h(x)$에서 α에 해당하는 값을 찾아야 하는데 이를 위해 사다리꼴 적분(trapezoidal rule) 방법을 사용하였다(Burden and Faires, 2000). 사다리꼴 적분 방법의 기본적인 아이디어는 적분구간을 다수의 사다리꼴로 나눈 후 이들의 면적을 계산하고 이들의 합을 이용하여 적분값을 구하는 것이다. 본 논문에서는 사다리꼴의 개수로 충분히 큰 수인 1,000을 사용하였다.

3.2 붓스트랩을 활용한 임계값 결정

붓스트랩은 확률분포를 가정하지 않고 실제 결과 값을 리샘플링(resampling)하여 분포를 추정하는 비모추 추정 방법이다(Efron and Tibshirani, 1993).

붓스트랩 방법을 사용하여 *k*NNDD의 임계값을 결정하는 과정은 다음과 같다. 먼저 아래에 보이는 바와 같이 *N*개의 정상 관측치의 *k*NNDD 값을 계산하여

$$kNNDD_1,\ kNNDD_2,\ \cdots,\ kNNDD_N$$

이를 B번 복원 추출하여 나열하면 아래와 같은 B개의 붓스트랩 샘플 셋을 얻을 수 있다. 여기서 붓스트랩 셋의 수인 B는 경우에 따라 다르지만 최소한 1,000번 이상을 사용하도록 권하고 있다(Efron and Tibshirani, 1993).

$$\begin{array}{cccc} kNNDD_1^{(1)}, & kNNDD_2^{(1)}, & \cdots, & kNNDD_N^{(1)} \\ kNNDD_1^{(2)}, & kNNDD_2^{(2)}, & \cdots, & kNNDD_N^{(2)} \\ \vdots & \vdots & \cdots & \vdots \\ kNNDD_1^{(B)}, & kNNDD_2^{(B)}, & \cdots, & kNNDD_N^{(B)} \end{array}$$

구해진 B개의 붓스트랩 샘플 셋 각각에서 사용자가 정한 0과 1사이의 α값을 통해 $100 \cdot (1-\alpha)$분위수를 구하고 이를 D_i $i = 1, 2, \cdots, B$라 하면 최종적으로 임계값은 D값의 평균을 취함으로써 구할 수 있다. 이를 표현하면 식 (6)와 같다.

$$\text{임계값} = \frac{1}{B}\sum_{i=1}^{B} D_i \tag{6}$$

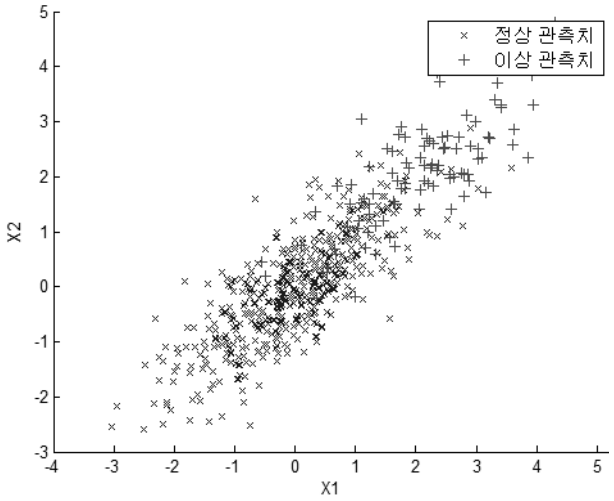
이를 이용하여 새로운 관측치의 *k*NNDD 값이 임계값보다 크면 이상치로 판단하게 된다.

4. 실험설계

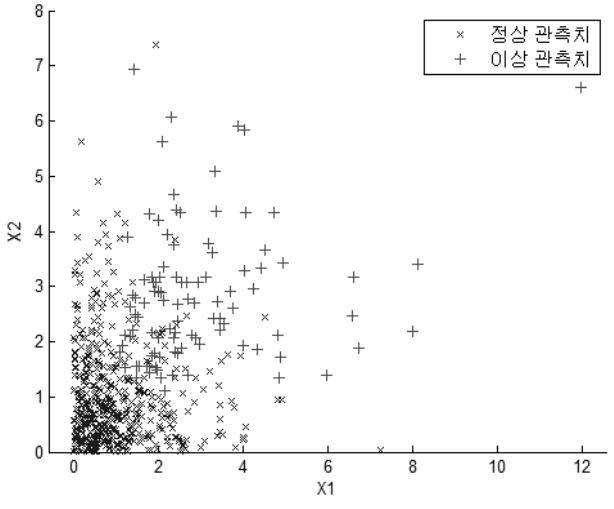
4.1 데이터

이번 장에서는 제안한 임계값 결정 방법의 특성과 성능을 살펴보기 위해 세 가지 다른 분포를(정규분포, 감마분포, 스위스 롤 분포) 통해 생성된 시뮬레이션 데이터를 사용하였다. <Figure 2>는 각각의 분포로부터 생성된 정상 관측치와 이로부터 평균이 이동된 이상 관측치를 보여주고 있다.

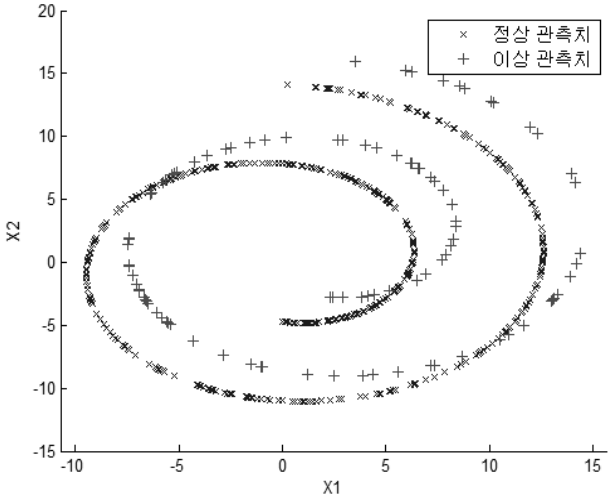
시뮬레이션은 Phase I과 Phase Ⅱ의 두 단계로 나누어 시행하였는데 Phase I은 정상 데이터를 이용해 임계값을 결정하는 단계이고 Phase Ⅱ는 Phase I에서 정해진 임계값을 통해 새로운 관측치가 정상인지 이상인지를 판단하는 단계이다. Phase I을 위해서 정상 관측치 500개를 생성하였고 Phase Ⅱ를 위해서는 정상 관측치 100개, 평균을 변화(mean shift)시킨 이상 관측치 100개를 생성하였다.



(a)



(b)



(c)

Figure 2. Simulation data (a) normal distributed data, (b) gamma distributed data, and (c) swiss roll data

4.2 실험결과

본 실험에서는 실험의 복잡도를 줄이기 위해서 kNNDD 거리 계산 시 동일한 값의 k_1과 k_2를 사용하였고 그 값으로는 100과 200, 두 가지 경우에 대해 살펴보았다.

실험결과의 비교를 위해서는 평가지표로써 널리 쓰이고 있는 민감도와 특이도를 사용하였다. 민감도는 전체 이상 데이터를 얼마나 정확하게 이상 데이터로 판단했는지를 나타내는 지표로써 아래와 같이 나타낼 수 있다.

$$Sensitivity = \frac{TP}{TP+FN} \tag{7}$$

즉, 전체 이상 데이터 중 올바르게 이상으로 판단한 비율을 의미한다.

특이도는 정상 데이터를 얼마나 정확히 정상으로 예측했는지를 나타내는 평가지표로써 다음의 식을 통해 계산할 수 있다.

$$Specificity = \frac{TN}{TN+FP} \tag{8}$$

즉, 총 정상 데이터 중 올바르게 정상으로 판단한 비율을 나타낸다. 식 (7)과 식 (8)에서 사용된 TP, FN, TN, FP에 대한 정의는 <Table 2>에서 살펴볼 수 있다.

시뮬레이션 데이터를 통해서 기존 임계값과 제안한 방법의 임계값을 비교하였다. <Table 1>은 기존 임계값을 정규, 감마, 스위스 롤 데이터에 적용하여 얻는 민감도와 특이도 결과를 보여주고 있으며 <Figure 3>, <Figure 4>, 그리고 <Figure 5>는 제안한 비모수 추정 방법을 각각 세 가지 분포로부터 생성된 데이터에 적용하여 얻은 민감도와 특이도 결과를 보여주고 있다. <Table 1>에서는 임계값을 일률적으로 1의 값을 사용한 기존의

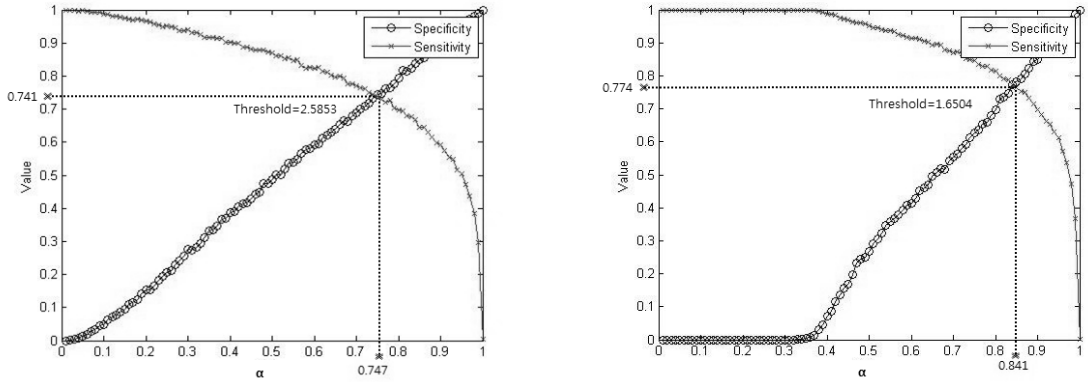
Table 1. Sensitivity and specificity of normal, gamma, and swiss roll distributed data with threshold = 1 in an *k*nndd algorithm

		$k_1 = k_2 = 100$	$k_1 = k_2 = 200$
Sensitivity	Normal Data	0.999	0.966
	Gamma Data	1.000	1.000
	Swiss Roll Data	0.778	0.569
Specificity	Normal Data	0.239	0.479
	Gamma Data	0.272	0.268
	Swiss Roll Data	0.298	0.395

Table 2. Confusion Matrix

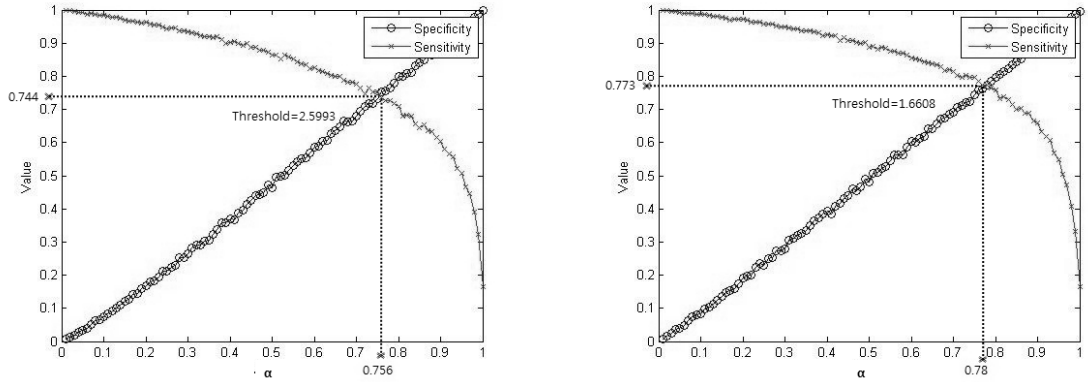
		Predicted	
		이상 관측치	정상 관측치
Actual	이상 관측치	TP (True Positive)	FP (False Positive)
	정상 관측치	FN (False Negative)	TN (True Negative)

임계값 결정 방식의 결과이다. 예를 들어, $k_1 = k_2 = 100$일 경우 기존 임계값을 정규분포에 적용해 보면 민감도와 특이도가 각각 0.999와 0.239임을 알 수 있다. 이 경우 민감도는 매우 높게 나왔으나 반대로 특이도는 매우 낮게 나옴을 알 수 있으며 이는 임계값으로 사용한 1이 적절한 선택이 아님을 보여주고 있다. 같은 경우 제안한 방법을 사용한 결과는 <Figure 3>에서 볼 수 있듯이 α에 따라 다양하게 얻어지는 민감도와 특이도 정보를 이용하여 사용자가 임계값을 유연하게 결정할 수 있음을 보여주고 있다. 특히, 민감도와 특이도가 서로 만나는 점은 두 지표가 모두 최대가 되는 지점이고 이 경우 해당하는 임계값을 사용하면 민감도와 특이도 모두가 최대가 됨을 알 수 있다. <Figure 3>에서 $k_1 = k_2 = 100$일 민감도와 특이도가 만나는 지점의 값은 KDE의 경우 0.741, 붓스트랩의 경우에는 0.744가 나옴을 알 수 있으며 해당 임계값은 각각 2.5853과 2.5993이다. 이는 임계값으로 1을 사용한 기존 방법을 통해 구한 민감도 = 0.999, 특이도 = 0.239의 평균 값인 0.619보다 높음을 알 수 있다. <Figure 4>와 <Figure 5>는 제안하는 방법을 각각 감마분포와 스위스롤 분포로부터 생성된 데이터에 적용한 결과를 보여주고 있으며 이 결과 역시 기존의 방법에 비해 보다 유연하고 정확한 결과를 보여주고 있다.



K=100 K=200

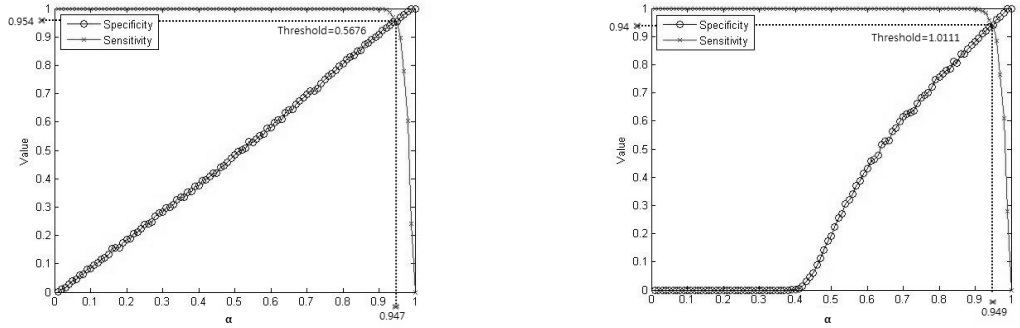
(a)



K=100 K=200

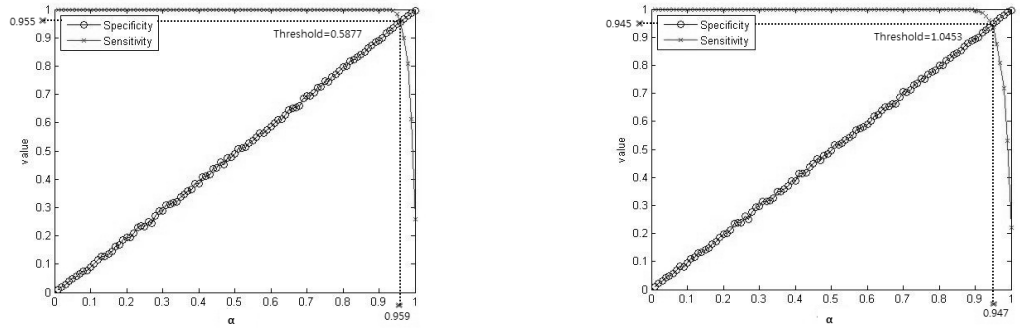
(b)

Figure 3. Results of normal distributed data by using (a) KDE and (b) bootstrap



K=100 K=200

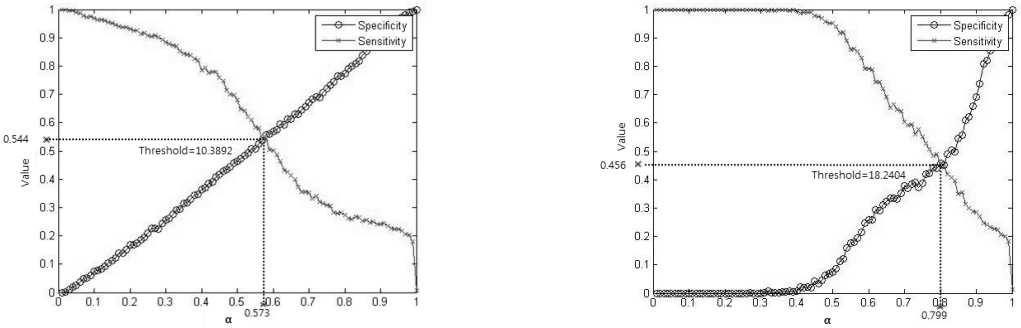
(a)



K=100 K=200

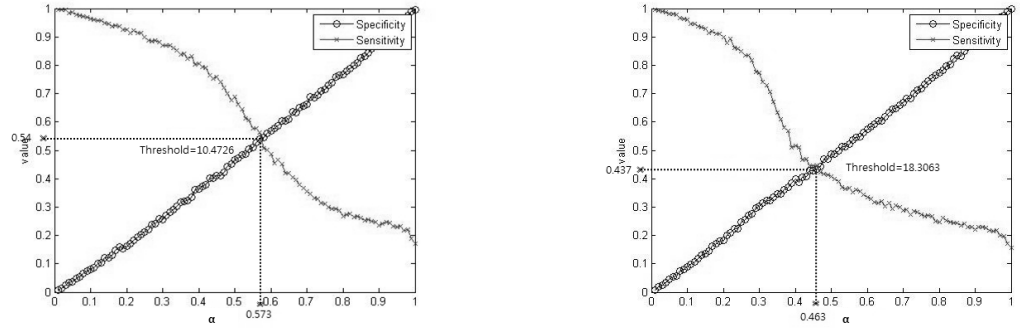
(b)

Figure 4. Results of gamma distributed data by using (a) KDE and (b) bootstrap



K=100 K=200

(a)



K=100 K=200

(b)

Figure 5. Results of swiss role data by using (a) KDE and (b) bootstrap

요약해 보면 제안한 임계값 결정 방법은 분석자의 의도를 반영할 수 있을 뿐만 아니라 비모수 기법을 사용하기 때문에 특정 분포에 대한 사전 정보가 필요가 없다는 장점이 있다.

5. 결 론

본 연구에서는 대표적인 비모수 추정 방법인 KDE 기법과 붓스트랩을 활용하여 단일 클래스 분류 기법인 *k*NNDD의 임계값을 결정하는 방법에 대해 제안하였다. 정규분포, 감마분포, 스위스롤 분포로부터 생성한 시뮬레이션 데이터를 통해 제안한 방법과 기존 방법을 비교해 제안한 방법의 성능을 검증하였다.

본 연구에서 제안하는 방법은 임계값 결정시 일률적으로 1을 사용하는 기존 방법과는 달리 민감도와 특이도의 정보를 이용하여 사용자가 유연하게 결정할 수 있는 장점이 있다.

본 연구는 단일 클래스 분류 기법중 하나인 *k*NNDD의 경우를 통해 새로운 임계값 결정 방법을 적용했지만 본 논문에서 제시하는 방법은 다른 단일 클래스 분류 기법의 임계값 결정문제에도 효과적으로 적용될 수 있을 것이다.

참고문헌

Alireza, Tavakkoli., Amol, Ambardekar., Mircea, Nicolescu., and Sushil, J. Louiset. (2007), A Genetic Approach to Training Support Vector Data Descriptors for Background Modeling in Video Data, *International Symposium on Visual Computing*, Lake Tahoe, 318-327.

Berry, M. J. A. and Linoff, G. (1997), Data Mining Techniques, John Wiley and Sons, Inc.

Breunig, M. M., Kriegel, H. P., Ng, R. T., and Sander, J. (2000), LOF : Identifying density-based local outliers. *in Proceedings of the ACM SIGMOD 2000 international conference on management of data*, **29**, 93-104.

Burden, R. L. and Faires, J. D. (2000), Numerical Analysis, Seventh Edition, Brooks/Core, Parcific Grove, CA.

Duda, R. and Hart, P. (1973), *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York.

Efron, B. and Tibshirani, R. (1993), An Introduction to the Bootstrap, Chapman and Hall/CRC, Boca Raton, FL.

Hand, David., Mannila, Heikki., and Smyth, Padhraic. (2001), Principles of data mining, Adaptive Computation and Machine Learning Series, MIT Press.

Khan, S. S. and Madden, M. G. (2010), A survey of recent trends in one class classication, *Articial Intelligence and Cognitive Science*-20th Irish Conference, Lecture Notes in Computer Science, **6206**, 188-197, Springer.

Kim, S. B., Sukchotrat, T., and Park, S. K. (2011), A nonparametric fault isolation approach through one-class classification algorithms, *IIE Transactions,* **43**, 505-517.

Koppel, M. and Schler, J. (2004), Authorship verification as a one-class classification problem, *in Proceedings of 21st International Conference on Machine Learning*.

Mason, R. L. and Young, J. C. (2002), Multivariate Statistical Process Control With Industrial Applications, *American Statistical Association and the Society for Industrial and Applied Mathematics*, Philadelphia, PA.

Manevitz, L. M. and Yousef, M. (2001), One-class svms for document classification, *Journal of Machine Learning Research*, **2**, 139-154.

Sanchez-Yanez, R. E., Kurmyshev, E. V., and Fernandex, A. (2003), One-class texture classifier in the CCR feature space, *Pattern Recognition Letter*, **24**.

Runger, G. C., Alt, F. B., and Montgomery, D. C. (1996), "Contributors to a multivariate statistical process control chart signal," *Communications in Statistics : Theory and Methods*, **25**(10), 2203-2213.

Silverman, B. W. (1986), Density Estimation for Statistics and Data Analysis. Chapmand and Hall, London, United Kingdom.

Song, S. I., Cho, Y. C., and Park, H. K. (2003), "Robust Control Chart using Bootstrap Method," *Journal of the Society of Korea Industrial and Systems Engineering,* **26**(3), 39-49.

Sukchotrat, T., Kim, S. B., and Tsung, F. (2010), "One-class classification-based control charts for multivariate process monitoring," *IIE Transactions*, **42**, 107-120.

Tan, P. N., Stein, M., and Kumar, V. (2007), Introduction to data-mining, Infinity books, Seoul, Korea.

Tax, D. M. J. (2001), One-class classification : Concept-learning in the absence of counter-examples, PHD thesis, Delf University of Technology, Netherlands.

Wasserman, Larry. (2006), All of nonparametric statistics, Springer Texts in Statistics.

Woodall, W. H. and Montgomery, D. C. (1999), "Research issues and ideas in statistical process control," *Journal of Quality Technology,* **31**(4), 376-386.